SCORE FUNCTIONS FOR ASSESSING CONSERVATION IN LOCALLY ALIGNED REGIONS OF DNA FROM TWO SPECIES UCSC TECH REPORT UCSC-CRL-02-30

KRISHNA M. ROSKIN, MARK DIEKHANS, W. JAMES KENT, AND DAVID HAUSSLER

ABSTRACT. We construct several score functions for use in locating unusually conserved regions in genome-wide search of aligned DNA from two species. We test these functions on regions of the human genome aligned to mouse. These score functions are derived from properties of neutrally evolving sites on the mouse and human genome, and can be adjusted to the local background rate of conservation. The aim of these functions is to identify regions of the human genome that are conserved by evolutionary selection, because they have an important function, rather than by chance. We use them to get a very rough estimate of the amount of DNA in the human genome that is under selection.

CONTENTS

Funding for this project was provided by NHGRI Grant 1P41HG02371. We thank Simon Whelan, Nick Goldman, Laura Elnitski, Ross Hardison, Webb Miller, Scott Schwartz, Francesca Chiaromonte, Aran Smit, Eric Lander, Bob Waterston and Francis Collins for their input and data.

1 INTRODUCTION

As part of the Mouse Genome Project, groups at several universities have been studying alignments between the draft genomes of human and mouse. A full report will be submitted by the Mouse Genome Sequencing Consortium at a later time. Here we report some preliminary results we obtained using early versions of this data[1]. We designed several score functions, described below, that could be applied to short aligned regions (tens to thousands of bases) to measure how diverged they were between the two species. Emphasis was on counting the number of observed base substitutions in various ways, although gaps are also considered in some versions of the score functions. We have been especially interested in looking at the distributions of these score functions on regions of aligned DNA that we have reason to believe are not under selection, but rather are evolving neutrally. We looked at two types of "neutral" sites:

(1) 4d-sites: 3rd bases in the 8 four-fold degenerate codons (sites marked "x" in the codons GCx (ALA), CCx (PRO), TCx (SER), ACx (THR), CGx (ARG), GGx (GLY), CTx (LEU), GTx (VAL) that can be any base without changing the amino acid)

(2) AR-sites: "ancient repeat" sites from retrotransposons or DNA transposons that were inserted in the genome before the human-mouse split and appear in syntenic positions in both species.

The properties of these sites will be described more fully in subsequent papers. In particular, we noticed that substitutions at a given site are dependent on the flanking bases, so some of our score functions take this into consideration. We hope to give a more complete treatment of this subject in a future paper as well. Here we use information from our study of neutrally evolving sites to construct some simple score functions for human-mouse aligned regions.

The score functions are:

- normalized divergence (Section 2)
- I-score (Section 3)
- context-dependent I-score (Sections 4 and 6)
- context-dependent I-score with gap penalties (Sections 5)

We first define these functions for gap-less aligned regions only, then we discuss ways of extending them to include gap costs. In the initial results below, to apply the score function to a gapped alignment, we just remove the gaps and indels first (see example in Section 5 below).

In the final section we use one of our score functions (the context-dependent I-score) to get a crude estimate of the fraction of the human genome that is under selection. To do this, we scored all non-overlapping 100bp windows with at least 30 aligned bases in the human genome draft, and plotted the empirical distribution of the scores we obtained (see Figure 13). We noticed an extra mass in the region where the scores for more highly conserved windows lie. This extra mass is absent when we plot the distribution of the scores from only the windows from ancient repeats, which are our model for typical scores from neutrally evolving DNA (see bell-shaped curve in Figure 13 representing the score distribution for windows of neutrally evolving DNA). We suspect this extra mass represents windows containing DNA that is under selection. Indeed, windows containing coding exons and known regulatory elements (kindly provided by Laura Elinski at Penn State University) do tend to have scores in the range where we see this extra mass in the genome-wide score

[1]The Mouse data was taken from the October Phusion assembly and aligned to the UCSC August Golden Path assembly using BLAT [9]. This data can found at `http://genome.ucsc.edu/cgi-bin/hgGateway?db=mm1`.

distribution (Figure 12). We obtain a crude estimate of the size of this extra mass by simply scaling the curve in Figure 13 for the density of the neutral distribution to fit within the overall density for the genome-wide scores, using the value at the origin. The neutral density is symmetric about this value, and the fit to the genome-wide density for all windows is quite good on the side representing scores from highly diverged regions, nearly all of which are likely to be neutral. On the side of more highly conserved regions, this scaling of the neutral density leaves out the extra mass that is likely to represent windows that are under selection. By subtracting the two densities, we find that this "selected" mass represents about 5% of the human genome.

It is clear that considerable further work needs to be done to validate and improve this estimate, including more sophisticated analysis of the densities, better and more complete assemblies of the genomes of both species, more exploration of the sensitivity of the method to the choice of windows and score functions, and a better understanding of the properties of neutrally evolving DNA, so that it may be more precisely distinguished from DNA under selection. Extending these methods to alignments of multiple species would sharpen the results considerably as well. We suspect that this will be required to reliably distinguish selected regions from neutral regions on a window-by-window bases, rather than in a "bulk statistical estimate" as we attempt here.

2 DIVERGENCE

Let $A = (a_1, ..., a_n; b_1, ..., b_n)$ be a gap-less alignment where a_j is the human base aligned to the mouse base b_j. Define

$$X_j = \begin{cases} 0 & \text{if } a_j = b_j, \\ 1 & \text{if } a_j \neq b_j. \end{cases} \tag{2.1}$$

Let

$$X = \sum_{j=1}^{n} X_j. \tag{2.2}$$

Then $D = X/n$ is the observed divergence in the alignment, i.e. the fraction of positions where the bases differ.

The score D is highly dependent on the length of the alignment, making it difficult to compare scores from alignments of different lengths. We convert D to a normalized divergence score (a "Z-score") in the standard way. Let m be the fraction of bases that differ in a global "reference" set of alignments representing neutral evolution, e.g. all 4d-sites or all AR-sites. To define a normalized divergence score for the alignment A under a model for neutral evolution induced by this reference set, we assume that random variables $X_1, ..., X_n$ are independent, and that they have a common mean m, that is we assume the X_j are i.i.d. The normalized divergence score for the alignment A is then:

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - mn}{\sqrt{(1-m)mn}} \tag{2.3}$$

If the i.i.d. assumption is satisfied, then the random variable Z should, by the central limit theorem, be approximately normal for large enough n. (Already for $n \geq 20$ the fit is not too bad if m is not too close to 0 or 1.)

In real human-mouse alignment data, the empirical distribution of Z over the sets of alignments representing neutral evolution that we have examined has been far from normal, exhibiting a variance much larger than 1. Figure 1 is the empirical distribution of the variable Z for 4d sites from approximately 8000 pairs of orthologous genes between human and mouse. For each orthologous pair, we formed an alignment A as defined above consisting

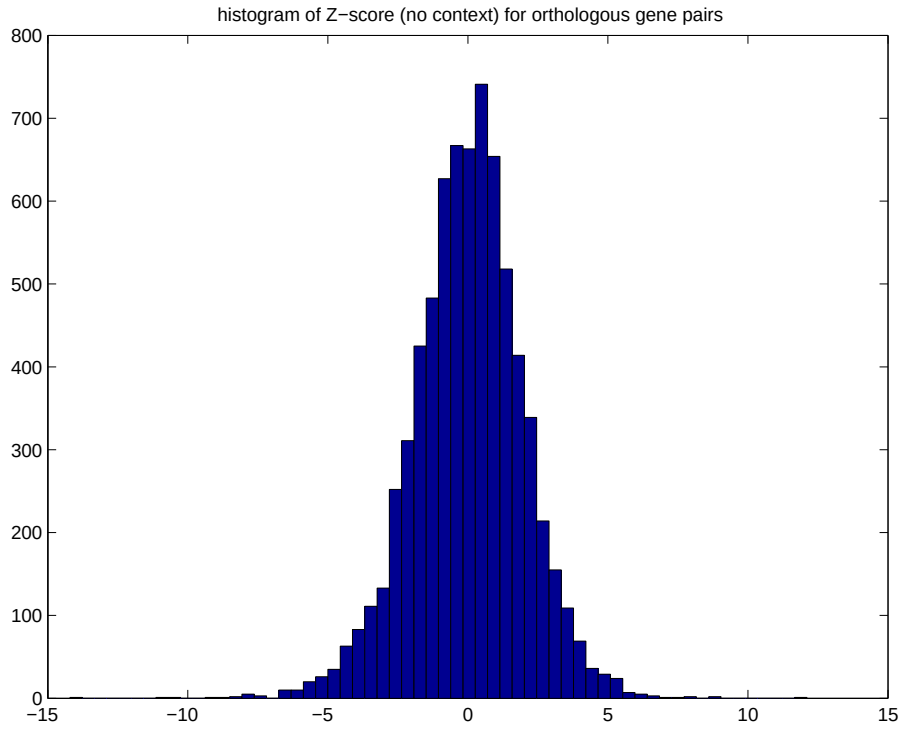

FIGURE 1. Histogram of the normalized non-context sensitive Z-score for the 4d data.

only of the 4d sites for this pair of genes. Figure 1 shows the histogram of this score for all pairs of genes with $n \geq 60$ 4d sites. The variance of this empirical distribution is 3.8696.

We repeated this analysis for the alignments of ancient repeats produced by Scott Schwartz and Webb Miller from Penn State University. We obtained a smaller variance of 2.4012 (see Figure 2, but still much too large for the score to be normal. The assumption that substitutions are i.i.d. is clearly rejected.

Before addressing the problem of modeling dependence between observed changes in an alignment so that we can obtain a properly normalized score, we first develop score functions that are based more directly on simple probability models of observed changes.

3 I-SCORE

One problem with the divergence score is that it treats equally all observed base changes, whereas in reality transitions are more than three times as frequent as transversions in the human-mouse data. In fact, all 16 possible observed changes (including the 4 identities) occur with different probabilities. It is customary to use loglikelihood ratios derived from these probabilities in constructing an alignment score function, so that each of the observed changes has its own "weight" in the overall score function[1, 5, 14].

Given a gap-less alignment $A = (a_1, \ldots, a_n; b_1, \ldots, b_n)$ as above, let

$$X_j = \log \frac{Q(a_j)R(b_j)}{P(a_j, b_j)} \tag{3.1}$$

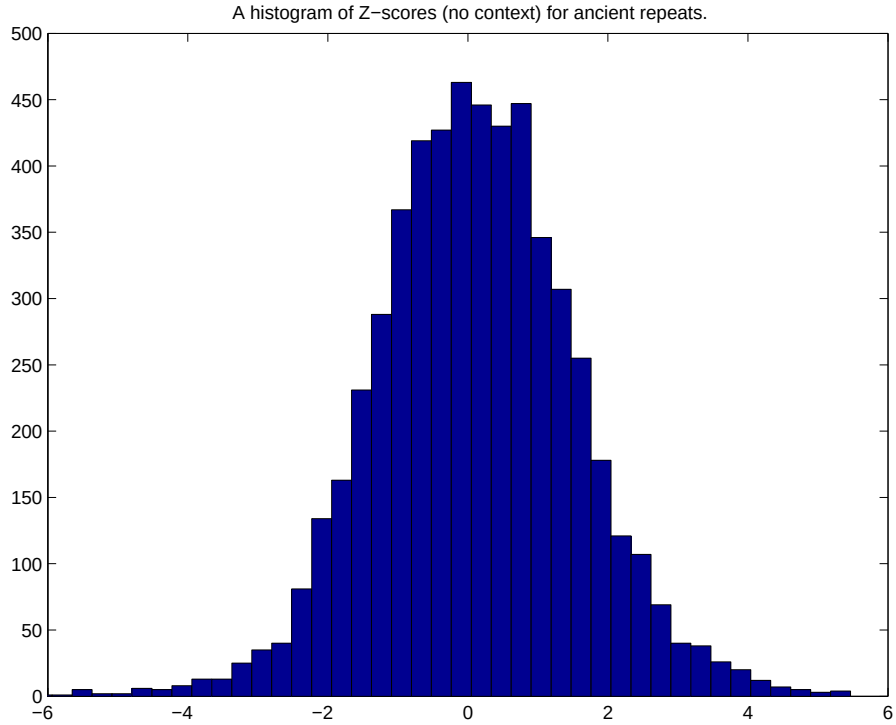

FIGURE 2. Histogram of the normalized non-context sensitive Z-score of ancient conserved repeats.

where $Q(a_j)$ is the fraction of times the base a_j occurs in the human sequences in the set of reference alignments ("human background probability" of a_j), $R(b_j)$ is the fraction of times the base b_j occurs in the mouse sequences ("mouse background probability" of b_j), and $P(a_j, b_j)$ is fraction of times the aligned pair (a_j, b_j) occurs in the reference alignments ("the paired background probability" of (a_j, b_j)).

Since $Q(a_j)$ and $R(b_j)$ are the marginals of $P(a_j, b_j)$, if we compute $-\mathrm{E}(X_j)$, the negative expectation of X_j with respect to $P(a_j, b_j)$, we get the mutual information between a_j and b_j [3]. Thus $-\mathrm{E}(X_j)$ is the average information that a mouse base gives about an aligned human base (and vice-versa, since mutual information is symmetric). Mutual information is always positive by Jensen's inequality[3]. It follows that $\mathrm{E}(X_j)$ is always negative for any probability distribution $P(a_j, b_j)$. Normally, X_j is negative when $a_j = b_j$ and positive otherwise, so it can also be easily viewed as a weighted measure of observed divergence.

Let A be the alignment of $a = a_1 \cdots a_n$ and $b = b_1 \cdots b_n$, and $X = \sum_j X_j$, where X_j is defined as in Equation 3.1 above. Then X has a simple probabilistic interpretation as well. Let M_0 denote the null hypothesis that the bases of a and b are independent and identically distributed according to the human and mouse background probabilities, respectively. Let M_1 denote the hypothesis that the aligned base pairs of a and b are independent, but within a pair, the two bases are dependent and distributed according to the paired background probabilities. Then it is easy to see that X is the loglikelihood ratio for the alignment A, given by

$$X = \log \frac{P(a, b|M_0)}{P(a, b|M_1)}. \tag{3.2}$$

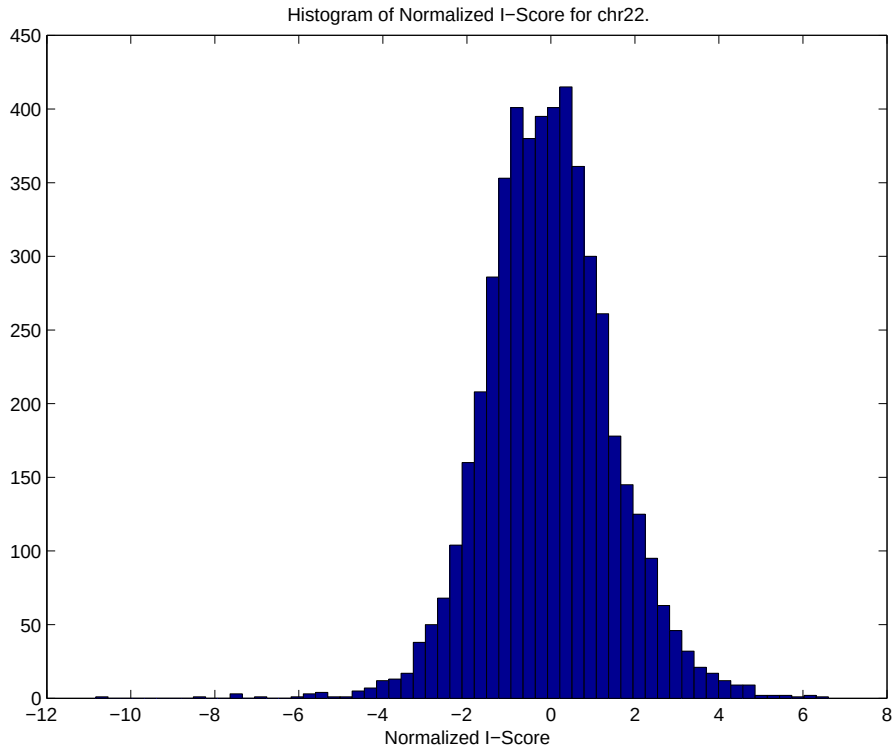

FIGURE 3. Histogram of I-Scores of ancient repeats on chromosome 22.

Define

$$\mathrm{E}(X) = \mathrm{E}(X|M_1), \tag{3.3}$$

the expectation with respect to model M_1. Here, in analogy with the properties of $\mathrm{E}(X_j)$ discussed above, since model M_0 is the product of the two marginal distributions of M_1 for sequences a and b, the quantity $-\mathrm{E}(X)$ is the mutual information between a and b, and hence $\mathrm{E}(X)$ is always negative.

If we continue with the above assumptions, then X is a sum of i.i.d. random variables, and thus we can reasonably define the Z-score from X as we did for the observed divergence in Equation 2.3 above,

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - \sum_j \mathrm{E}(X_j)}{\sqrt{\sum_j \mathrm{Var}(X_j)}}. \tag{3.4}$$

We call this the I-score of the alignment A. Since we are assuming that the X_j are i.i.d., the expressions $\mathrm{E}(X_j)$ and $\mathrm{Var}(X_j)$ are just constants independent of j that can be estimated from a global "reference" set of alignments representing neutral evolution, e.g. all 4d-sites or all AR-sites, as we did for the mean m in calculating the normalized divergence score.

A histogram of the I-score for all of the ancient repeats on human chromosome 22 is given in Figure 3. This dataset has a variance of 2.2427. A histogram of the I-score for the 4d sites of genes on human chromosome 22 is given in Figure 4. This has a variance of 3.9400. Again, the variances larger than 1 indicate that the assumption of independence of the observed changes, and hence independence of the X_j, is violated in actual alignments.

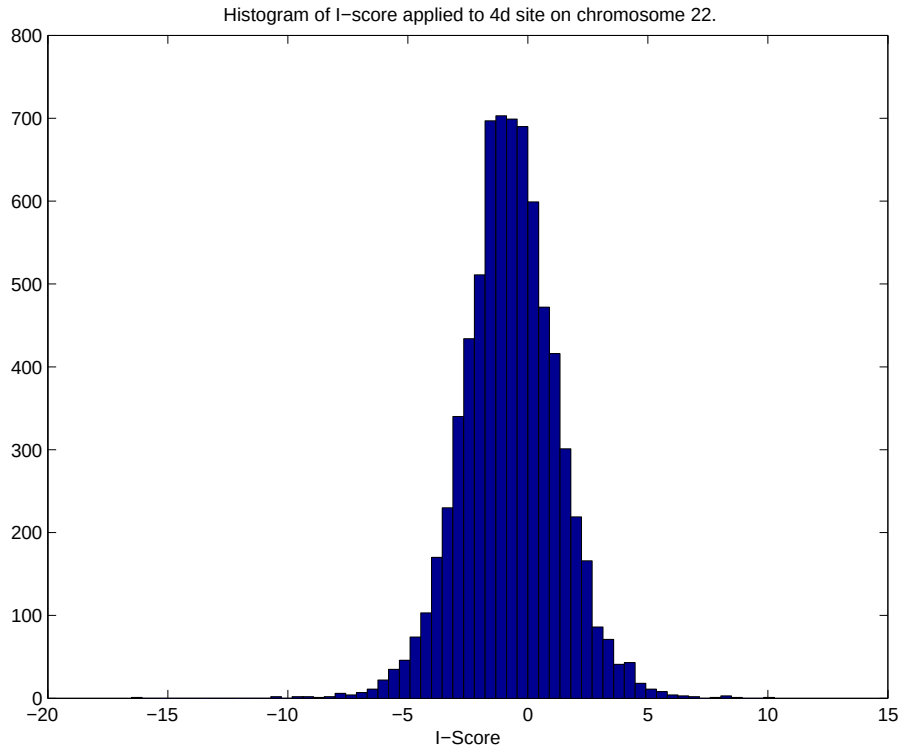

FIGURE 4. A histogram of the I-score applied to 4d site of genes on chromosome 22.

4 CONTEXT-DEPENDENT I-SCORE

One aspect of the dependence of the X_j in the above two score functions is the strong effect of flanking bases on observed base changes. This includes the "CpG" effect that has been heavily studied, and other strong observed effects. We can construct a version of the I-score based on a more realistic model of the probabilities of observed base changes probabilities by using context-dependent probabilities that take into account the flanking bases as "context" for the observed change.

Let $c_j = (a_{j-1}, b_{j-1}, a_{j+1}, b_{j+1})$ be the context of the aligned pair of bases (a_j, b_j) in the gap-less alignment A of $a = a_1 \cdots a_n$ and $b = b_1 \cdots b_n$. We assume aligned pairs of bases (a_0, b_0) and (a_{n+1}, b_{n+1}) are provided to handle the boundary conditions. To take the contexts into account in the I-score of alignment A, we can modify the I-score to use the conditional background probabilities $P(a_j, b_j|c_j)$, defining

$$X_j = \log \frac{Q(a_j|c_j)R(b_j|c_j)}{P(a_j, b_j|c_j)}, \tag{4.1}$$

where, as above, Q is the human marginal of P and R is the mouse marginal of P. These probabilities are obtained by counting the observed frequencies of the different observed changes separately in all possible contexts, using data collected from either 4d- or AR-sites.

We let $X = \sum_j X_j$, as in Equation 2.2 above, but replacing the definition of X_j by Equation 4.1, and define the context-dependent I-score as

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}},$$

in analogy with Equation 2.3 above.

To calculate $\mathrm{E}(X)$, we can still use the expansion $\mathrm{E}(X) = \sum_j \mathrm{E}(X_j)$. However, the X_j are no longer identically distributed. Rather, $E(X_j)$ depends on the context c_j. There are 256 possible values for c_j, depending on the flanking bases for site j in human and mouse. As we have many millions of aligned pairs of bases and their contexts in the data set of alignments from AR-sites, we can get good estimates for the 256 possible values of $\mathrm{E}(X_j)$ from this data, and simply use these in our calculation of $\mathrm{E}(X)$.

It is tempting to try the same thing for $\mathrm{Var}(X)$, expanding as $\mathrm{Var}(X) = \sum_j \mathrm{Var}(X_j)$ as above, but that would require that the X_j are assumed independent. This would not make sense, as their contexts overlap, and in addition, a similar assumption has seemed to get us into trouble in approximating the normalized divergence and I-score as well. Here is a simple alternate approximation.

Let X_{odd} be the sum of the X_j for odd j between 1 and n, and X_{even} be the sum of the X_j for even j. We make the decomposition

$$X = X_{\mathrm{odd}} + X_{\mathrm{even}} \tag{4.2}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(X_{\mathrm{odd}}) + \mathrm{Var}(X_{\mathrm{even}}) + 2\,\mathrm{Cov}(X_{\mathrm{odd}}, X_{\mathrm{even}}). \tag{4.3}$$

Then we make the approximations

$$\mathrm{Var}(X_{\mathrm{odd}}) = \sum_{j \text{ is odd}} \mathrm{Var}(X_j) \tag{4.4}$$

$$\mathrm{Var}(X_{\mathrm{even}}) = \sum_{j \text{ is even}} \mathrm{Var}(X_j) \tag{4.5}$$

and

$$\mathrm{Cov}(X_{\mathrm{odd}}, X_{\mathrm{even}}) = C_0 \sqrt{\mathrm{Var}(X_{\mathrm{odd}})\,\mathrm{Var}(X_{\mathrm{even}})} \tag{4.6}$$

where C_0 is an empirically estimated constant. Thus we approximate the context-dependent I-score as

$$Z = \frac{X_{\mathrm{odd}} - \mathrm{E}(X_{\mathrm{odd}}) + X_{\mathrm{even}} - \mathrm{E}(X_{\mathrm{even}})}{\sqrt{\mathrm{Var}(X_{\mathrm{odd}}) + \mathrm{Var}(X_{\mathrm{even}}) + 2C_0 \sqrt{\mathrm{Var}(X_{\mathrm{odd}})\,\mathrm{Var}(X_{\mathrm{even}})}}} \tag{4.7}$$

The justification for the approximations in Equation 4.4 and 4.5 is as follows.

Let C_{odd} be set of contexts for all the aligned pairs of bases in the terms of X_{odd}, and analogously for C_{even}. Note that since the aligned pairs at even indices serve as contexts for the aligned pairs at odd indices and vice-versa, C_{odd} is the set of aligned pairs of bases with even indices and C_{even} is the set of aligned pairs of bases with odd indices. In approximation Equation 4.4, we are assuming that the aligned pairs of bases at odd sites are conditionally independent, given C_{odd}, their contexts consisting of the aligned pairs of bases at even sites, and analogously for approximation Equation 4.5. This assumption of conditional independence given flanking context bases is milder than the assumption of strict independence, and factors out the immediate effects of the flanking bases. (Of course, we could do further expansion on $\mathrm{Var}(X)$ as well, perhaps including those terms in the standard quadratic expansion at distance 2, 3, 4, etc., but we leave that for further research.)

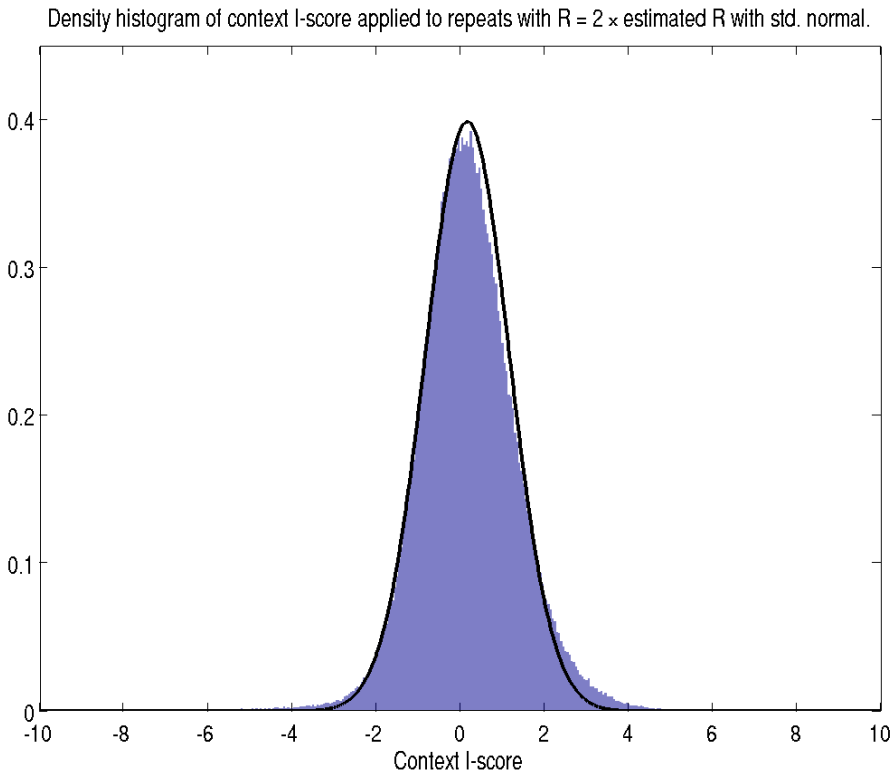

FIGURE 5. Histogram for context-dependent I-scores of the aligned ancient repeats on chromosome 22.

The justification for Equation 4.6 derives from the observation that the correlation coefficient[13] between X_{odd} and X_{even},

$$R = \frac{\text{Cov}(X_{\text{odd}}, X_{\text{even}})}{\sqrt{\text{Var}(X_{\text{odd}})\,\text{Var}(X_{\text{even}})}} \tag{4.8}$$

might be expected to scale roughly as a constant independent of the length n of the alignment. Hence, empirically estimating R and setting $C_0 = R$ in Equation 4.6 would be a reasonable way to estimate $\text{Cov}(X_{\text{odd}}, X_{\text{even}})$. In practice, with human-mouse alignments of AR-sites on chromosome 22, we find that $R = 0.34$, nearly independent of the length n, but that $C_0 = 2R$ empirically gives a more normal distribution for the approximate context-dependent I-score Z defined in Equation 4.7. Thus we set $C_0 = 0.68$.

The histogram for context-dependent I-scores of the aligned ancient repeats on chromosome 22 is given in Figure 5, with the normal distribution superimposed.

There is a fatter tail than expected on the right side, for highly diverged alignments, but the fit to the normal tail on the left side, for highly conserved alignments, is excellent. This is good, because we intend to use the score to find human-mouse alignments that are significantly more conserved than would be expected from the model of neutral evolution estimated from the alignments of ancient repeats.

Importantly, the distribution of the context-dependent I-score does not have a visible dependence on the length n of the alignment. In Figure 6 and Figure 7 and we show the histogram for context-dependent I-scores of the small aligned ancient repeats on chromosome 22 (length n of the aligned region between 50 and 62), and for comparison, the scores

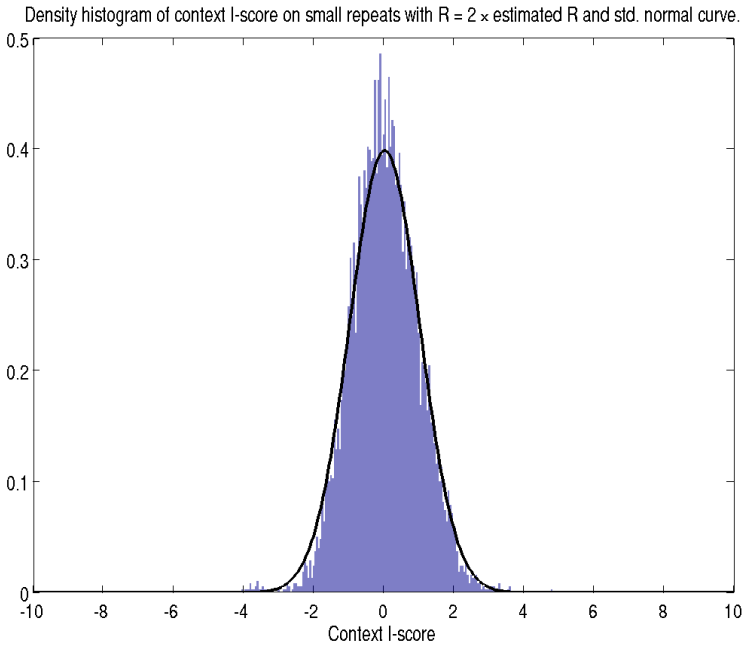

FIGURE 6. Histogram for context-dependent I-scores of the small aligned ancient repeats on chromosome 22.

for more than 10-fold longer alignments obtained by concatenating copies of the shorter alignments

5 INCLUDING INSERTIONS AND DELETIONS IN THE SCORE

Until this point, we have only been considering scores for ungapped alignments, simply removing the gaps from gapped alignments as necessary before calculating the score functions. However, the gaps themselves do contain information that can be included in the score function. For example, the gapped alignment

```
A--CTG----CCGATTGC
AGGCAGTTTT---AT--C
```

with human on top and mouse on the bottom, is reduced to the gap-less alignment

```
1234567
ACTGATC
ACAGATC
```

with sites numbered above it. The simple I-score from Section 3 is then derived from the probabilities of the 7 pairs of the aligned bases (A,A), (C,C), (T,A), etc., assuming independence. However, between each of these 7 consecutive pairs, we can use the gapped alignment to define an indel event. In particular, between sites 1 and 2 there is either an insertion of `GG` in mouse or a deletion of `GG` in human. For simplicity, let us treat all these indel events as insertions, denoting, e.g., an insertion of `GG` in mouse as `M:GG`, and a similar insertion in human as `H:GG`. Let us also refer to the case where there is no insertion in one species as a null insertion. Finally, let us simply ignore indels at the ends of the alignment, assuming these are trimmed off as is customary. Under these assumptions, along with a

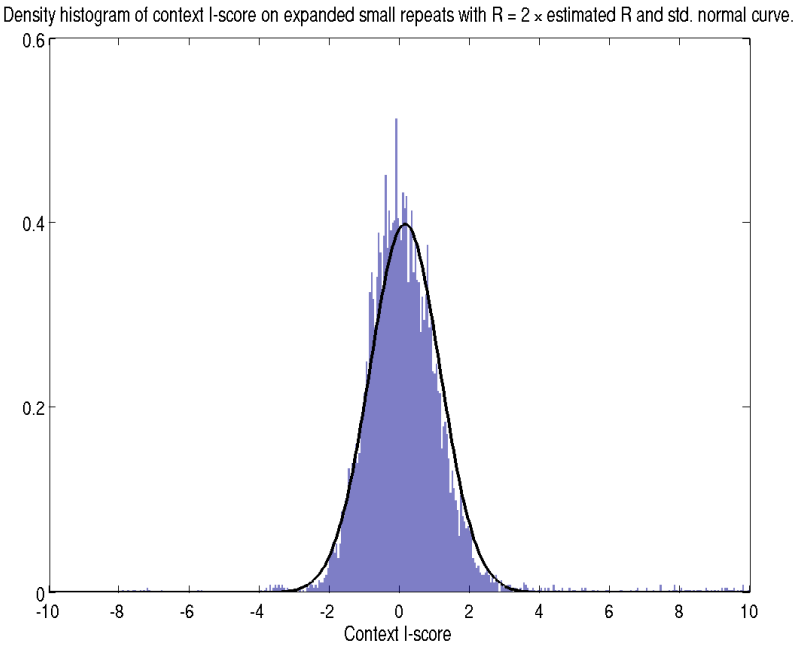

FIGURE 7. Histogram for context-dependent I-scores of the small aligned ancient repeats on chromosome 22 that have been expanded ten-fold.

reduced gap-less alignment of length n, we also obtain a list of pairs of insertion events of length $n-1$. E.g., for the above example, this list is

```
(H:null,M:GG), (H:null,M:null), (H:null,M:null),
(H:CCG,M:TTTT), (H:null,M:null), (H:TG,M:null).
```

In general, we will denote this list as

$$(r_1, s_1), (r_2, s_2), \ldots, (r_{n-1}, s_{n-1}).$$

The simplest score model for a gapped alignment treats each insertion event as independent. In analogy with the definition X_j for the I-score, let

$$Y_j = \log \frac{Q(r_j)R(s_j)}{P(r_j, s_j)} \tag{5.1}$$

where $Q(r_j)$ is the fraction of times the human insertion r_j occurs between two aligned positions in a reference set of gapped alignments ("human background probability" of r_j), $R(s_j)$ is the analogous thing for the mouse insertion s_j, "mouse background probability" of s_j), and $P(r_j, s_j)$ is fraction of times the insertion pair (s_j, r_j) occurs in the reference alignments ("the paired background probability" of (s_j, r_j)).

We then redefine X to include the log odds score for both the pairs of aligned bases and the pairs of insertions between them by letting

$$X = \sum_{j=1}^{n} X_j + \sum_{j=1}^{n-1} Y_j, \tag{5.2}$$

Then, as in Equation 3.4, we define the I-score with indels as

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \tag{5.3}$$

$$= \frac{X - \sum_{j=1}^{n} \mathrm{E}(X_j) - \sum_{j=1}^{n-1} \mathrm{E}(Y_j)}{\sqrt{\sum_j \mathrm{Var}(X_j) + \sum_j \mathrm{Var}(Y_j)}}. \tag{5.4}$$

In practice, it is not possible to empirically estimate $\mathrm{E}(Y_j)$ and $\mathrm{Var}(Y_j)$ for all possible pairs of insertions. We require a simplified model for $P(r_j, s_j)$.

Let us define $l(r_j)$ to be the length of the sequence that is inserted in the insertion r_j, and similarly for s_j. The length of a null insertion is 0. Similarly, define the sequence that is inserted in the insertion r_j as $S(r_j)$, and similarly for s_j. We may decompose the probability $P(r_j, s_j)$ as

$$P(r_j, s_j) = P[S(r_j), S(s_j)|l(r_j), l(s_j)]P[l(r_j), l(s_j)]$$

we can then make the assumption that the actual sequences that are inserted at a given position separately in the human and mouse lineages are independent, given their lengths. Thus

$$P(r_j, s_j) = P[S(r_j)|l(r_j)]P[S(s_j)|l(s_j)]P[l(r_j), l(s_j)]$$

Since Q and R are the marginals of P, making a similar decomposition yields

$$Q(r_j) = P[S(r_j)|l(r_j)]Q[l(r_j)]$$

and

$$R(r_j) = P[S(s_j)|l(s_j)]R[l(s_j)]$$

Thus

$$Y_j = \log \frac{Q[l(r_j)]R[l(s_j)]}{P[l(r_j), l(s_j)]} \tag{5.5}$$

i.e. Y_j does not depend on the sequences that are inserted between sites j and $j+1$, but only on the lengths of the inserts.

In practice there is an upper limit K on the size of insert that is allowed. If an alignment contains an insertion larger than this size it is broken into two alignments that are scored separately. In such a case we often have enough empirical data to estimate $P(n_1, n_2)$ for all insert lengths n_1 and n_2 between 0 and K, and form a table of observed frequencies for values of Y_j. This can be used to estimate $\mathrm{E}(Y_j)$ and $\mathrm{Var}(Y_j)$.

An alternative is to break the length distributions into a probability that the length is zero, and a (conditional) geometric length distribution given that the length is not zero. This leads to a variant of the well-known affine gap penalties used often used in scoring alignments[5]. The above probabilistic formulation reduces to the type of score function used for pair-HMMs in this case, which are probability models for gapped alignments[5].

An extreme case is to only distinguish null from non-null insertions. Let

$$p = P[l(r_j) > 0 \text{ and } l(s_j) > 0] \tag{5.6}$$

and

$$q = P[l(r_j) > 0 \text{ and } l(s_j) = 0] = P[l(r_j) = 0 \text{ and } l(s_j) > 0]. \tag{5.7}$$

Then there are only three cases for Y_j:

(1) If there is no insertion in either species between sites j and j+1 then

$$Y_j = \log \frac{(1-p-q)^2}{1-p-2q}.$$

(2) If there is an insertion in one species but not the other then

$$Y_j = \log \frac{(1-p-q)(q+p)}{q}.$$

(3) If there is an insertion in both species then

$$Y_j = \log \frac{(q+p)^2}{p}.$$

Note that if $p = 0$, i.e. there is never an insertion in both species, and q is small, i.e. insertions in either species are rare, then if we use natural logs, we can approximate case (1) by $Y_j = q^2$ and case (2) by $Y_j = -q$. Thus if there are k places out of $n-1$ where there is an insertion in the alignment, the total raw score X defined in Equation 5.3 above is approximately

$$X = \sum_{j=1}^{n} X_j - kq + (n-k-1)q^2 \tag{5.8}$$

A little algebra then shows that we can approximate Equation 5.3 by

$$Z = \frac{\sum_{j=1}^n X_j - \sum_{j=1}^n E(X_j) - kq(1+q^2) + 2(n-1)(1+q)q^2}{\sqrt{\sum_{j=1}^n \mathrm{Var}(X_j) + 2(n-1)(q^3)[(1-q-q^2)^2 + 2q(1-2q)(1+q)^2]}} \tag{5.9}$$

$$\sim \frac{\sum_{j=1}^n X_j - \sum_{j=1}^n E(X_j) - kq(1+q^2) + 2(n-1)(1+q)q^2}{\sqrt{\sum_{j=1}^n \mathrm{Var}(X_j) + 2(n-1)(q^3)}} \tag{5.10}$$

for small q. This gives a variant of the I-score introduced in Section 3 above with one additional parameter q that models indels. An analogous variant of the context-dependent I-score is also defined similarly. In both cases we are assuming that the indels are independent from each other, and from the aligned bases that flank them. Weaker assumptions are possible, but cumbersome.

6 FURTHER EXTENSIONS

It makes biological sense that the probability of seeing a particular observed base change (a_j, b_j) at an AR-site j would depend on more than the context c_j of the flanking bases in human and mouse, defined in the previous section. Indeed, we observe that context features like the percentage of G+C bases in a window of 20,000 bases around the human ancient repeat also significantly affect the probability of seeing particular observed change; other authors have previously reported this effect as well [11, 2]. There is also evidence of effects of unknown origin that cause local regions of a chromosome to be more conserved than the genome-wide average, or to be more diverged[7, 15, 8, 10]. Thus we expect that probability of seeing a base change at an AR-site j may depend on the average number of changes per site in a window surrounding ancient repeat containing site j.

The simplest way to modify the score function that we have defined above to be sensitive to local variations is to estimate the observed frequencies of base changes using reference alignments from a local window rather than genome-wide. This makes all of the expectations and variances in the above formulas depend on the local region of the genome you are analyzing. The difficulty is in the trade-off between choosing a large enough window to get good estimates and a small enough window to track variation in the parameters being estimated. Models with fewer parameters are favored since there are fewer parameters to estimate.

Regional variation creates correlations between the scores of nearby sites. Our data indicates that this correlation occurs at both large and small distances but occurs predominately over smaller distances. Figure 8 shows the squared correlation coefficient[13] (R^2) between the context-dependent I-scores for pairs of ancient repeat alignments with centers separated by various distances between 1 and 25,000 bases. Data is from selected chromosomes. A detailed plot for distances between 1 and 5,000 bases is given in Figure 9. The value plotted

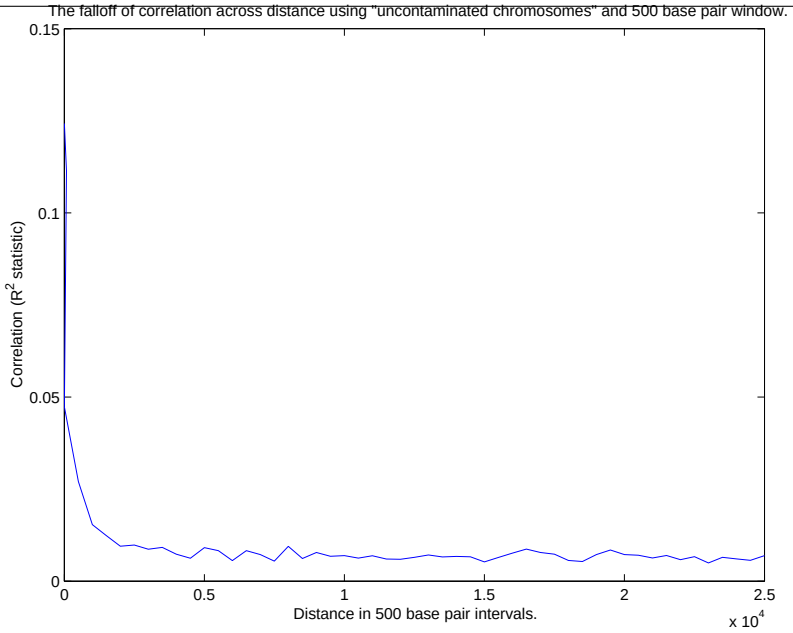

FIGURE 8. The falloff of the squared correlation coefficient (R^2) between the context-dependent I-scores for pairs of ancient repeat alignments with centers separated by various distances between 1 and 25,000 bases.

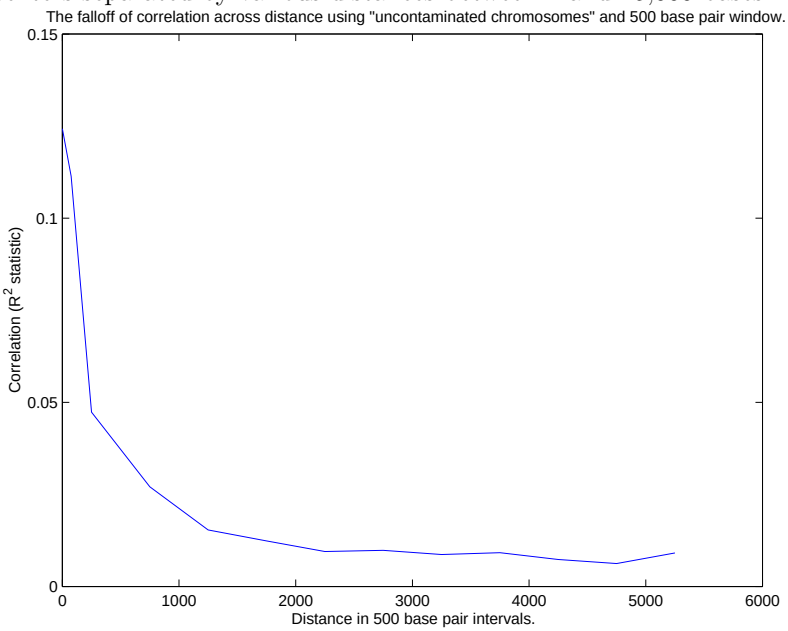

FIGURE 9. The falloff over distances between 1 and 5,000 bases.

for distance 1 is the square of the correlation coefficient between the scores for the observed changes in the odd-indexed sites vs. the even-indexed sites of a single ancient repeat alignment, defined in Equation 4.8 above. The value plotted at separation distance 90 bases is the analogous value for the correlation between the score of the first half vs. the second

half of the same ancient repeat alignment; since the typical alignment has about $n = 180$ aligned sites, the centers of the first and second halves are separated by about 90 bases. The remaining points, at separations of 250 bases, 750 bases, etc., are for correlations between the scores of two different repeats separated by roughly these distances.

Since R^2 measures the fraction of variance in one variable that is explained by the other, these results show that a significant amount of variance (more than 10%) in the context-dependent I-score of one half of an alignment is explained by the score of the other half. This is a type of dependence that is not accounted for directly in the approximation used in Section 4, and may account for the fact that C_0 needed to be bigger than R for the score to be approximately normal. This additional dependence falls off quickly, but is still fairly strong at distances less than 1000 bases. Only between 1 and 2 percent of the variance is explained by the context dependent I-scores of repeats separated by more than 1500 bases, and this effect persists for about 25,000 bases, only gradually falling at distances of more than several hundred thousand bases. This could very well be the effect of variation in G+C content over large regions[11], or other factors that cause variation in mutations rates over large regions. If we are able to further modify the context-dependent I-score to take into account the effects of G+C content and other factors affecting conservation in a window surrounding the alignment, it should make this score easier to normalize, and make it more useful for discriminating neutrally evolving regions for regions under selective pressure.

Let us represent the pair of aligned bases at a given site by a categorical random variable Y that takes on a value in the set 1, ..., 16. To model these more general types of effects on Y, we can use a generalized linear model (GLM) with response variable Y and a stimulus vector V with components measuring the percentage of G+C bases in a window of 20,000 bases and the average number of changes per site in say, 50, 1,000, and 20,000 base windows. The GLM is defined by

$$U = f(V) \tag{6.1}$$

where $U = U_1...U_{16}$ is a real vector and f is a linear function and

$$P(Y = i|V) = \frac{\exp(U_i)}{\exp(U_1) + \cdots + \exp(U_{16})} \tag{6.2}$$

The parameters of the GLM are the coefficients of the linear function f. For the (simple) I-score from Section 3 above, these can be estimated by pooling the genome-wide data from all AR-sites and using maximum likelihood.

This gives an alternate way of extending the I-score in Section 3 to include context dependence by replacing $P(a_j, b_j)$ with the function computed in Equation 6.2 in the definition of X_j from Equation 3.1. Here the index i is the one representing the pair of bases (a_j, b_j), and the marginal probabilities for the individual human and mouse bases, $Q(a_j)$ and $R(b_j)$ are calculated as the marginals of P.

In principle we could add 256 more features to the stimulus vector V to obtain a generalization of the context-dependent I-score from Section 4 such that the context c_j included not only the flanking bases for site j, but the G+C content and average number of changes in a surrounding window. However, in practice, it is safer to estimate 256 separate GLMs, one for each of the possible flanking bases, since we have enough data in the genome-wide human-mouse alignments of ancient repeats to do this accurately with maximum likelihood. We hope to explore this line of research in future work.

7 Tests of the Selected Score Functions

The goal of the context I-score with gap penalty is to identify abnormally conserved regions on human genome, regions that have to be conserved because they perform some key biological function.

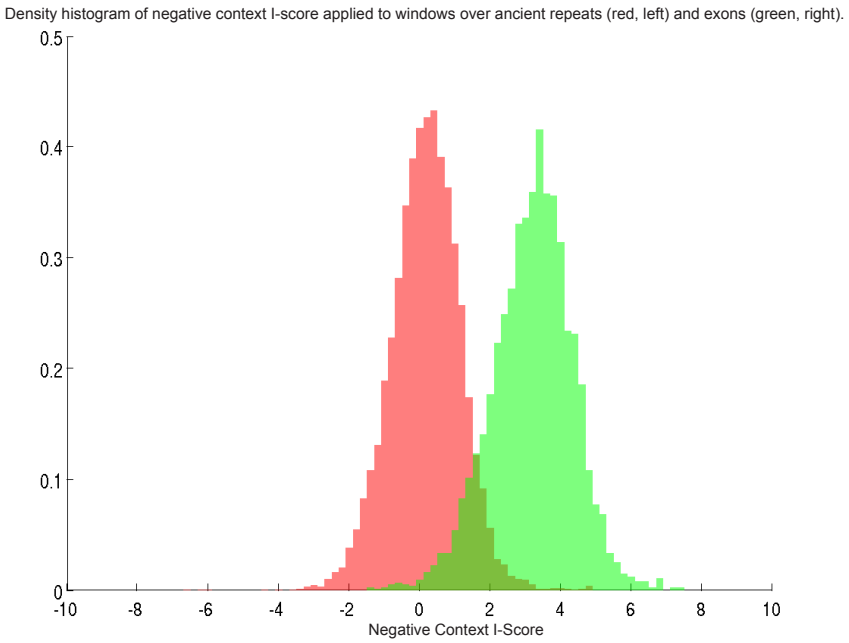

FIGURE 10. A histogram on the density scale of the values of the negative context I-score with gap penalty of windows over coding exons and ancient repeats. The dark green is where the scores overlap.

The mouse assembly was aligned to the human genome with BLAT[9]. The mouse/human alignment was then broken up into regions containing no inserts or deletions of length greater than six. Each of these regions was broken into window of between 30 and 100 bases that could overlap by as many as 50 bases. This produced a data set of about thirty-five thousand windows. This formed our central data set.

To estimate the parameters of the context I-score with gap penalty score function, we collected background probabilities from the window contained inside ancient repeats. This data was used to estimated the $Q(a_j|c_j)$, $R(b_j|c_j)$, and $P(a_j, b_j|c_j)$ probabilities for Equation 4.1, the covariance constant C_0 for Equation 4.7, the gap frequency parameter q of Equation 5.7. These parameters tune the score function to model "neutral" evolution.

One test of the performance of the score function was to see how well it can distinguish known functional regions from known non-functional regions. For our functional regions, we used coding exons in known genes and known regulatory elements. Gene data was taken from refSeq[12]. Regulatory element data was compiled by Laura Elnitski at Penn State from various sources. We call a window "over exons" if at least half of their bases come from exons. We call a window "over repeats" if the entire window contains bases from an ancient repeat. Likewise, we call window "over regulatory regions" if the whole windows is in a regulatory region. For these results, the ancient repeats and coding exons data is taken from chromosome 22. Regulatory regions are genome-wide.

To evaluate the performance, we scored windows over the functional regions and non-functional regions, as defined above. Figure 10 shows a density histogram of the scores of windows over exons and over ancient repeats. Here and in the remaining figures in the paper we plot the negative context I-score, which is a measure of conservation rather than divergence. Figure 11 shows a similar histogram for windows over regulatory regions and repeats.

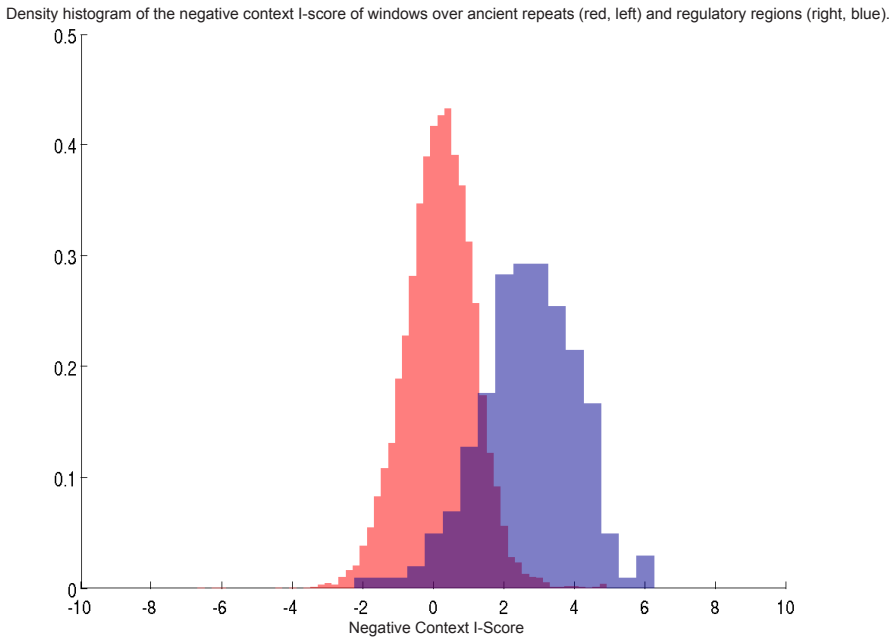

FIGURE 11. A histogram on the density scale of the values of the negative context I-score with gap penalty of windows over regulatory regions and ancient repeats. The dark purple is where the scores overlap.

We also scored all 100-base windows on human chromosome 22. The results are shown in in Figure 12. The red histogram on the left shows scores for the negative context I-score with gap penalty applied to windows over all of chromosome 22. The green histogram is the negative context I-score applied to know regulatory regions and coding exons on chromosome 22.

To try to recognize functional regions from non-functional regions using this score, we can define a threshold such that any windows with a score over that threshold we will call functional and anything below non-functional. To find the best possible threshold, we searched for the minimum error rate decision boundary[4]. We defined a loss of 1 for each functional region that was predicted to be non-functional and vice versa and a loss of 0 for a correct prediction. More complicated loss structures are possible. For instance, if the score function was going to be applied genome-wide, it would be better to have a more stringent threshold to prevent the flood of false positives. These other loss structures are not explored here. We used 3,996 functional windows and 79,920 non-functional windows; the 20-to-1 ratio reflects the 5% of the genome that we estimated to be functional (see below). The minimum error rate decision boundary was found to be at a negative context I-score of 2.46802. The error rate or misclassification rate with that boundary was 0.0233567. In other words, if you called everything in this dataset with with a score greater than 2.46802 functional and the rest non-functional, you would be right about 97.5% of the time. Of course, there are serious limitations to extrapolating this result to the genome as a whole. One would have to work with separate train and test sets, and to make sure that the test sets were truly representative of neutral and functional DNA in general. Using ancient repeats as a model for neutral evolution, there is always the danger of learning to discriminate DNA in ancient repeats from all the other DNA, rather than learning to recognize DNA under selection!

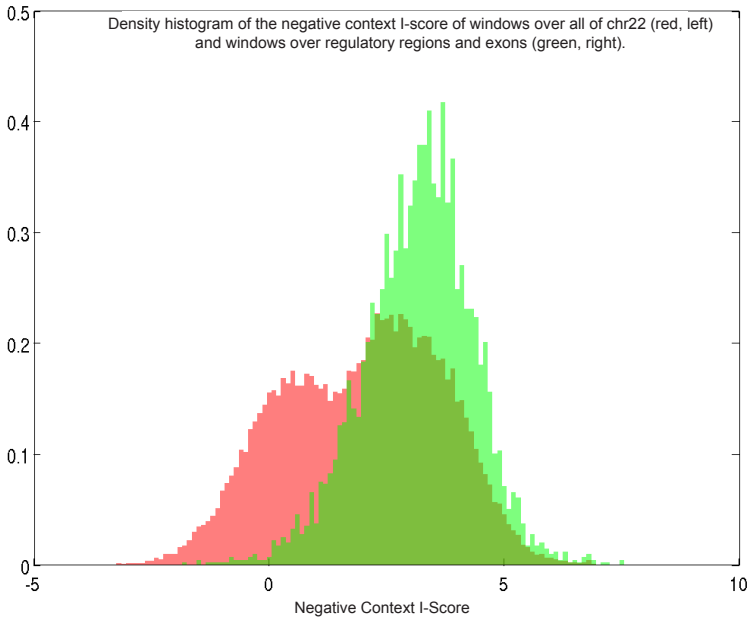

FIGURE 12. A histogram on the density scale of the negative context I scores of windows over all of chromosome 22 and a histogram of the scores of windows over know functional regions (coding exons and regulatory elements).

8 ESTIMATING THE FRACTION OF THE HUMAN GENOME UNDER SELECTION

There are roughly 40,000 windows from chromosome 22 that are scored in Figure 12. Since, windows can overlap by 50 bases, 40,000 windows is going to cover between 2 million and 4 million bases on chromosome 22, which is 10% of chromosome 22. The distribution of scores looks bimodal, with about half of the windows in each mode. If we call the higher scoring mode functional, we see that 5% of chromosome 22 is predicted to be functional by this measure.

To get a more accurate number for the fraction of the human genome under selection, we extended this calculation to use 100bp windows over the whole genome, not just chromosome 22, and use a more quantitative method. Figure 13 shows the distribution of the negative context-dependent I-scores with gap penalty of these windows. Note the disproportional number of high scoring windows that account for the right-hand bump. Compare this distribution of all windows to the distribution the scores of ancient repeat windows shown in the red histogram of Figure 14. We suspect that the right-hand bump in Figure 13 comes from the scores of windows that contain selected DNA, such as the windows over coding exons shown in the green histogram of Figure 14. If we call the distribution of ancient repeat scores neutral, we can estimate the fraction of all windows that do not belong to the neutral distribution and are thus under selection. The mean of the neutral distribution is $\mu = 2.4 \times 10^{-5}$, essentially 0. Since the neutral distribution is symmetric about its mean, the observed frequency of windows that score below the mean is 0.5. If we assume that the scores of windows that are under selection are positive, we can estimate the fraction of all windows that are neutral by looking at the percentage of 100bp windows over the whole genome that score below the neutral mean μ. This fraction was found to be 0.4078. Thus

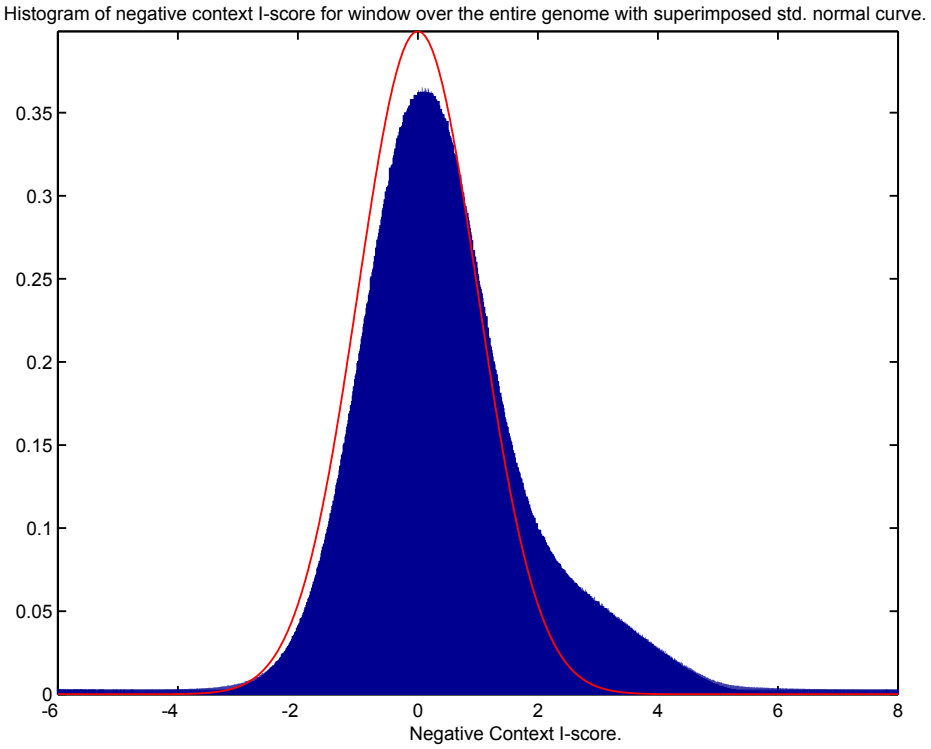

FIGURE 13. Histogram for the negative context-dependent I-scores with gap penalty of all windows genome-wide.

$2 \times 0.4078 = 0.8156$ or 81.56% of the windows in Figure 13 are from neutral DNA. It follows that $1 - 0.8156 = 0.1844$ of the windows are likely to be under selection. This fraction only takes into account human DNA that was alignable to mouse DNA. The number of 100bp windows used in the above calculations was approximately 8.4 million which cover about $100 \times 8.4 \times 10^6 = 8.4 \times 10^8$ of the 2.82 billion bases in the human genome. Assuming that the non-alignable sequence is too diverged to be under selection, we can estimate that

$$\frac{8.4 \times 10^8}{2.82 \times 10^9} \times 0.1844 = 0.055$$

or 5.5% of the human genome in 100bp windows that are under selection. This number is far grather than the 1.5% that is thought to be coding[6]. This may in fact be an under estimate because, as we can see in Figure 14, the scores of many windows in coding exons (which are likely to be in regions under selection) score less than μ. This means that the above calculation is in some sense conservative, and may be calling neutral many windows that are not. However, as discussed in the introduction, considerable further research will be required to determine the sensitivity of this method to assumptions, improve the density analysis, and look at other alignments, species and score functions, to fully validate this approach.

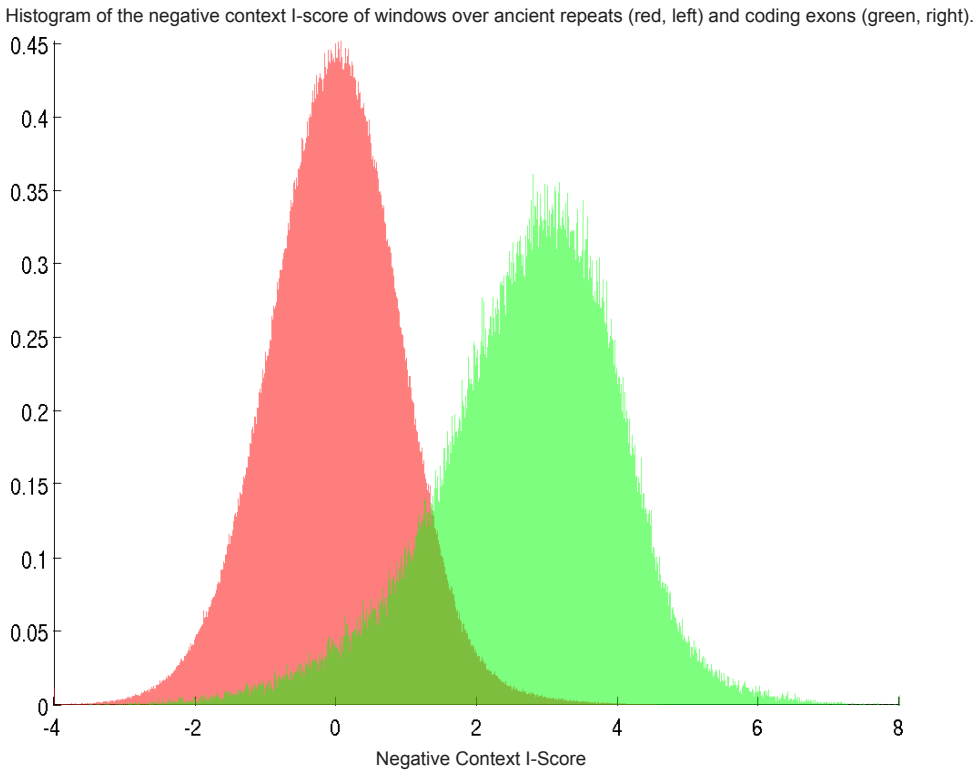

FIGURE 14. Histogram for the negative context-dependent I-scores with gap penalty of 100bp windows genome-wide in ancient repeat (red) and coding exons (green).

9 CONCLUSION

In this paper we have constructed score functions that are tuned to detect abnormally conserved regions on the human genome. By looking at conserved regions of the human genome we can predict key functional elements using these score functions, although the accuracy of the method is not fully determined. We have used the method to give a crude estimate of the fraction of the human genome under selection. Section 6 discusses a framework for adding new attributes that can further help determine the biological importance of a sequence. This may improve the score function. As the genomes of new species are sequenced, it may be worthwhile to generalize these scores to utilize the much greater information that will be contained in multiple genome alignments.

10 ACKNOWLEDGMENTS

The authors would like to thank: Scott Schwartz for aligning the ancient repeats in human to mouse, Webb Miller for sharing his findings on context-dependent rates of substitution, Arian Smit for identifying repeats that are shared by human and mouse, Laura Elnitski for collecting data on known regulatory elements. The authors would also like to thank Ross Hardison, Nick Goldman and Simon Whelan for their help.

REFERENCES

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool*, J. Mol. Biol. **215** (1990), 403–410.
2. R. D. Blake, S. T. Hess, and J. Nicholson-Tuell, *The influence of nearest neighbors on the rate and pattern of spontaneous point mutations*, J. Mol. Evol. **34** (1992), 189–200.
3. Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley, 1991.
4. Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern classification*, 2nd ed., Wiley-Interscience, October 2000.
5. Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.
6. Eric Lander et al, *Initial sequencing and analysis of the human genome*, Nature **409** (2001), 860–921.
7. Joseph Felsenstein and Gary A. Churchill, *A hidden markov model approach to variation among sites in rate of evolution*, Mol. Biol. Evol. **13** (1996), no. 1, 93–104.
8. J. Huelsenbeck and B. Rannala, *Phylogenetic methods come of age: testing hypotheses in an evolutionary context*, Science **276** (1997), 227–231.
9. W. James Kent, *The BLAST-like alignment tool*, Genome Research (2002).
10. G. Matissi, P. M. Sharp, and C. Gautier, *Chromosomal location effects of gene evolution in mammals*, Current Biology **9** (1999), 786–791.
11. Brian R. Morton, *The influence of neighboring base composition on substitutions in plant chloroplast coding sequences*, Mol. Biol. Evol. **14** (1997), no. 2, 189–194.
12. K. D. Pruitt and D. R. Maglott, *RefSeq and LocusLink: NCBI gene-centered resources*, Nucleic Acids Research **29** (2001), no. 1, 137–140.
13. John A. Rice, *Mathematical statistics and data analysis*, 2nd ed., Duxbury Press, June 1994.
14. D. Weaver, C. Workman, and G. Stormo, *Modeling regulatory networks with weight matrices*, 1999.
15. Z. Yang, *Among-site variation and its impact on phylogenetic analysis*, Tree **11** (1996), no. 9, 367–371.

CENTER FOR BIOMOLECULAR SCIENCE AND ENGINEERING AND HOWARD HUGHES MEDICAL INSTITUTE (D.H.) UNIVERSITY OF CALIFORNIA–SANTA CRUZ, SANTA CRUZ, CA, USA

E-mail address: krish@soe.ucsc.edu

URL: http://www.soe.ucsc.edu/~krish/